DOCUMENT RESUME

ED 164 572 TM 007 845

AUTHOR Bell, Robert M.
TITLE Medical School and Physician Performance: Predicting Scores On The American Board of Internal Medicine Written Examination.
INSTITUTION Rand Corp., Santa Monica, Calif.
SPONS AGENCY Health Resources Administration (DHEW/PHS), Bethesda, Md. Div. of Medicine.
REPORT NO Rand-R-1723-HEW
PUB DATE Aug 77
CONTRACT NO1-MB-24196
NOTE 43p.
AVAILABLE FROM Rand Corporation, 1700 Main Street, Santa Monica, California 90406

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Graduate Medical Education; *Graduate Medical Students; Higher Education; *Medical Schools; *Occupational Tests; Performance Factors; *Physicians; *Predictor Variables; *Success Factors
IDENTIFIERS *American Board of Internal Medicine Examination; National Board Examinations

ABSTRACT

Scores from the American Board of Internal Medicine certification examination (ABIM) were analyzed to relate the quality of physician performance (best indicated by ABIM scores) to early factors in the physician's training. Data were obtained on 438 1955, 1960, 1965, and 1969 graduates of nine unnamed medical schools. Regression analyses were performed to determine the importance of various stages of the training period on the ABIM score. Significant differences were found between the nine schools; a substantial amount of this difference was due to the quality of students admitted. Preadmission variables, however, accounted for less than 14% of the variance in ABIM scores. Medical school class rank accounted for 22%, and National Board scores--taken in the junior or senior year of medical school were even better predictors. Much more of the separation in scores was attributable to the medical school period than to preadmission variables. No significant relationship was found between ABIM scores and postgraduate training. It was found, however, that graduates taking the examination as soon as possible after eligibility performed better than those who waited. The measurement error, 11 to 25% of the total variance, was fairly small considering the homogeneity of the sample. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED164572

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MEDICAL SCHOOL AND PHYSICIAN PERFORMANCE: PREDICTING SCORES ON THE AMERICAN BOARD OF INTERNAL MEDICINE WRITTEN EXAMINATION

PREPARED FOR THE HEALTH RESOURCES ADMINISTRATION
DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

ROBERT M. BELL

R-1723-HEW
August 1977

TM007 845

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Rand Corp
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

2

Rand

The work upon which this publication is based was performed pursuant to Contract No. NO1-MB-24196 with the Health Resources Administration, Department of Health, Education, and Welfare. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the Department of Health, Education, and Welfare.

Published by The Rand Corporation

1700 Main Street, Santa Monica, Ca. 90406 Phone (213) 393-0411

Principal Investigator: Albert P. Williams

Contract Project Officer: Joseph Millard Brown
Division of Medicine
Public Health Services
Health Resources Administration
Department of Health, Education, and Welfare
Hyattsville, Maryland 20782

PREFACE

This report documents research performed under contract with the Bureau of Health Resources Development of the Health Resources Administration, and the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health, Education, and Welfare. The work is part of a larger Rand study of the effects of federal programs on academic health centers (Contract No. NO1-MB-24196).

The academic health center is an organizational complex that includes a medical school, at least one and usually several teaching hospitals, and often semi-autonomous research institutes as well as other health professional schools. The federal effects on the centers stem from the activities of a dozen federal agencies administering well over 100 distinct programs. This study addresses the quality of physicians at the end of their postgraduate training. The data for the study were scores on the written specialty certification examination given by the American Board of Internal Medicine.

SUMMARY

This study is an attempt to relate the quality of physician performance to early factors in the physician's training. Doing such a study is made difficult by the near impossibility of finding a satisfactory measure of so hazily defined a concept. The best available measures of performance seem to be scores on the specialty certification examinations. These examinations test the physician's comprehension of a large amount of pertinent information--usually within a few years of medical school graduation. Although these examinations certainly have limitations, they do measure a significant aspect of the physician's ability to give good care.

The research described here is part of a larger study on the effects of federal programs on academic health centers. During the course of that study data have been collected on students from a sample of representative U.S. medical schools. This includes only general information about the physicians once they have left medical school--type and length of postgraduate training, specialty, type of practice, etc. To obtain a later measure of performance, The Rand Corporation contacted the American Board of Internal Medicine for results on their written examination. The ABIM certifies more physicians than any other American specialty board. It provided Rand with the results of all specialty and subspecialty board examinations administered by them to 1955, 1960, 1965 and 1969 graduates of nine medical schools, through the year 1973. The sample includes 438 students--1955, 1960, 1965, and 1969 graduates of nine U.S. medical schools. A summary of the larger study will be found in Albert P. Williams et al., *The Federal Government and Academic Medicine: The Effects of Federal Programs on Activities and Outputs*, The Rand Corporation, R-1814-HEW (forthcoming). The report also details the selection of the academic health centers used in the study.

The ABIM written examinations are standardized each year to have a mean score of 500 points and a standard deviation of approximately 100 points within a reference group of recent U.S. or Canadian medical school

graduates taking the exam for the first time. This study determines the contribution of various factors to the variance in exam scores. Least squares linear regressions were performed with the examination score as the dependent variable. Various factors characterizing the physician's background and preparation were used as prediction variables. The results of these regressions permit inferences to be made about the relative importance of various stages of the training period on the exam score.

The major question considered is how much of a difference the medical school makes in the scores. Quite significant differences were found in scores among the nine schools, the difference between the largest and smallest estimated school effect being 55 points. A substantial amount of this difference is due to the quality of students admitted by the institutions. When preadmission variables are taken into consideration, the magnitude of the estimated school effects drops significantly. In fact, the possibility remains that if better measures of preadmission performance were available, it might turn out that which school a physician attended made no difference at all. For this sample of schools, a high correlation was found between estimated school effects and a composite measure of research activity at the institutions. This finding for a small number of schools does not prove that there is any true relation, but it does suggest a circular relationship in which the same schools tend to do more research, attract better students, produce better physicians, and gain more prestige. This is just what makes it difficult to separate the effects of each factor on the others.

Although preadmission variables and medical school attended definitely do have some effect, the ability to predict ABIM test scores at the beginning of medical school is slight. For example, preadmission variables account for less than 14 percent of the variance. The single variable rank within medical school class accounts for 22 percent. When they are available, National Board scores are even more useful predictors, the best ones being the internal medicine section and the mean score on Part II, covering the clinical sciences, both of which are taken in the junior or senior year of medical school. All in all, much more of the separation in scores is attributable to the medical school period than to the preadmission period.

No relationship was found between ABIM exam scores and postgraduate training. None of the available variables do a good job of differentiating among programs. The only distinctions are such crude ones as the type of hospital at which training was done (major teaching or limited teaching) and the type of internship (specialized or rotating). Further, the quality of postgraduate training is highly dependent on previous achievement, which makes it difficult to separate the effects of the two periods.

Before taking the written examination a physician must complete certain training requirements. Over the years, graduates who have gained board eligibility and have taken the exam at the earliest possible time outperformed those who have had some delay. This is to be expected since better graduates are more likely to be admitted immediately into preferred training programs. However, even when all other variables are used to account for achievement during medical school, the length of the delay is still quite significant. The decrease in predicted score due to each year of delay ranges from about 10 to 25 points, depending on the amount of information already in the regression. In general, the less information on previous achievement available, the more useful the delay period is. The fact that this variable is so significant suggests that there are systematic differences that are not accounted for by the other variables but that are somehow discernible to the candidates themselves and to the medicine program directors who influence when the examination is taken.

Every examination of this nature has a measurement error caused by the fact that a given candidate usually would not get the same score on two parallel administrations of the test. The measurement error represents variance that the background variables are not expected to explain. To assess the quality of fit of the regressions accurately requires knowledge of the size of the error variance. It is estimated that over the period in question, the average error variance was between 11 and 25 percent of the total variance. Considering the rather homogeneous nature of the examination group, this is fairly small. Even so, when it is taken into account, the most complete regression accounts for about

55 to 65 percent of the explainable variance. Thus, the ABIM written examination score correlates well with other evaluation criteria.

ACKNOWLEDGMENTS

The author is grateful to the American Board of Internal Medicine for supplying the results of its written examination. Extensive care was taken by both the Board and the author to ensure the confidentiality of the information in this and the other files. At no time were the names of physicians linked with their scores on the written examination. George Webster and John Meskauskas of the Board helped to answer questions about the administration of the examinations.

Thanks are also due to the administrators of the sample medical schools for releasing and helping gather data on their students. The staff of the Center for Health Services Research and Development of the American Medical Association provided postgraduate training information on these students from the American Medical Association Masterfile of Physicians. Kent Brown of Rand assisted in the preparation of some of the files.

Albert Williams and John Koehler were instrumental in initiating the study and contributed helpful comments and guidance. Finally, the author is grateful to David Armor, John Rolph, and Gus Haggstrom for criticisms of earlier drafts of the report. He is, of course, solely responsible for any errors that remain.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS ix

Section

I INTRODUCTION 1

II FACTORS AFFECTING THE ABIM WRITTEN EXAMINATION SCORE 4

Explanatory Variables 4
Treatment of Missing Data 5
Hypothesis Testing Using Hierarchical Stepwise Regression 6
Rank in Class and Age As Explanatory Variables 7
Test for School Effects 10
Introduction of National Board Scores . . . 12
Postgraduate Training Variables 13
Relationship of Score with Delay in Taking Test 14

III MEASUREMENT ERROR IN THE ABIM WRITTEN EXAMINATION 18

IV CONCLUSIONS 21

APPENDIX

USING REPEATED TAKINGS TO ESTIMATE THE MEASUREMENT ERROR OF THE WRITTEN EXAMINATION 26

I. INTRODUCTION

The last few years have seen a great variety of studies of U.S. medical schools and physicians, but few have been comprehensive analyses of physician performance in relation to the medical school experience. There are several difficulties in attempting such an analysis. First, there is no consensus on how quality of physician performance should be measured, so there are no standard criteria accepted by the medical community. Studies that have chosen some criteria have usually been limited to a single region and have lacked many details in the physicians' histories.

An appropriate surrogate for physician performance appears to be the scores on qualifying examinations of the various American specialty boards. These examinations are painstakingly created by top physicians throughout the country. Additionally, since the examinations are usually taken within a few years of medical school graduation, results for recent classes can be obtained. Until this present study, however, no one has had use of the data to do any study of these tests.

A written examination is given annually by the American Board of Internal Medicine (ABIM) to physicians who have met minimum post-graduate training requirements. Currently, three years in approved programs in internal medicine, one of which may be the internship year, are needed for qualification. No written examination was administered in 1971, at which time major revisions were made in the certification procedure. Before that time both a one-day written examination and a subsequent bedside oral examination were required. Beginning in 1972, a two-day written examination replaced these two. Besides multiple choice and true-false questions, the test contains a substantial number of "patient management problems." These simulated case histories require the physician to make diagnostic decisions and specify treatments based on the information received during the course of the problem.

While the oral examination was given, only a pass or fail mark was reported. For the written examination, however, a numerical score is

given. The numerical score is standardized each year within a reference group of U.S. and Canadian medical school graduates, taking the test for the first time, who completed training within one year of examination. The distribution of scores in the reference group is close to a normal distribution with a mean of 500 points and a standard deviation of approximately 100 points. The minimum passing score has varied from year to year. The data used in the analysis include scores for 438 physicians who had taken the written examination one to five times.[1] A few physicians took the written examination a second time after once passing it, as an alternative to taking the oral examination during its phasing out. Only scores of first attempts are used in the main part of the analysis.

This study focuses on determining which early factors in a physician's career influence the quality of his work once he is in practice. We are interested in such questions as: Can good physicians be distinguished from poor ones when they are admitted to medical school? Is there a significant difference among U.S. medical schools in the quality of their graduates? And how important is the postgraduate training period in the process?

The scores on the ABIM examination are not necessarily synonymous with the ultimate quality of a physician's practice. A test cannot measure a physician's dedication to his profession, his inclination to keep abreast of new research, or his willingness to answer emergencies at 3:00 a.m. At best such a test measures only his knowledge of some of the tools required to provide good medical care.

The method of analysis is an ordinary least squares linear regression with the written examination score as the dependent variable. The independent variables used as predictors are introduced hierarchically to facilitate the testing of certain hypotheses. Section II describes the regression procedure in greater detail, defines the independent variables used, and explains how missing values were handled. That section also contains the regression results.

There is a certain amount of measurement error inherent to the examination. That is, a person's score is not predetermined by the knowledge he takes into the testing room. Knowledge of the size of this error is important in assessing how well the regressions predict

the examination scores. Section III contains an assessment of the size of that error. The mathematical basis for that discussion is developed in the appendix.

Section IV contains a nontechnical summary of the regression results and the conclusions of the study.

FOOTNOTES TO SECTION I

[1]The American Board of Internal Medicine provided us with the results, through 1973, of all specialty and subspecialty board examinations administered to the 1955, 1960, 1965, and 1969 graduates of nine medical schools. Background data on these students were obtained directly from the records of the individual medical schools and from the AMA's *Master File of Physicians*.

II. FACTORS AFFECTING THE ABIM WRITTEN EXAMINATION SCORE

A number of least squares regressions have been performed with the score on the first taking of the ABIM written examination as the dependent variable. In addition to regressions on the entire sample of 438 students, some regressions were done for each year separately to check for inconsistencies and changes over time.

EXPLANATORY VARIABLES

Explanatory variables from three time periods were used. Pre-medical school variables, obtained from admission files, include undergraduate grade point average,[1] a selectivity index (1 lowest--9 highest) characterizing the undergraduate college, and scores from the four sections of the Medical College Admissions Test (MCAT). The indicators of performance at medical school are rank in class and the individual test scores and mean scores on Parts I and II of exams administered by the National Board of Medical Examiners. Since the pairwise correlations among the medical school performance indicators are high, only three National Board scores were used. The first is the overall mean of Part I on the basic sciences test, usually taken in the second year of medical school. Also selected were the overall mean and the internal medicine section of Part II on the clinical sciences test, usually taken in the last year. The National Board standardizes scores in the same way as the ABIM--to a mean of 500 and standard deviation of 100 for each taking. Each physician's classroom performance during the four years of medical school was measured by the variable rank in class. This variable has been transformed so that within each entire class it has close to a normal distribution with mean 500 and standard deviation 100.

The measures of postgraduate training are restricted to three indicator variables.[2] The first equals one if the student took a specialized internship and zero otherwise. The second variable indicates those students who had an internship at a major teaching hospital. The third variable indicates those students who had a residency or fellowship at a major teaching hospital *without* having had an internship at one. The last two variables cover nearly all combinations, since only four

students failed to follow major teaching hospital internships with a major teaching hospital residency or fellowship. Two took residencies at community hospitals and two showed no further training. The file also contains the total number of months spent in residency and fellowships, but since there are minimum requirements for admission to the examination, this variable has little meaning. In many cases it would only measure how long it took until the student was able to pass the test. In fact, for those students who attempted the examination, there is a nonsignificant negative correlation between the score on the written examination and total time in residency.

The medical school each student attended, the year of graduation, his age at graduation, and the number of years afterward that he took the test are also known.

TREATMENT OF MISSING DATA

Central to all of the regression work is the treatment of missing data values. Careful efforts have been made to avoid reporting biased or misleading results and still not lose power by reducing sample sizes. About 8 percent of the sample was missing one or more of the preadmission variables. Those students were left out of any regressions using that set of variables (other than age).

The only other variables with missing values were rank in class and the two parts of the National Board Examinations. The patterns of missing values of these variables make the simple treatment used for the preadmission variables undesirable. No data on rank in class were available at all from one medical school or for the 1969 graduating class at another. To remove those observations would have made it impossible to estimate all of the school coefficients while controlling for rank. There are similar circumstances for the National Boards, since almost all of a class usually took a particular part, or none did. In most cases the decision to take either or both parts was apparently an institutional rather than an individual decision. Thus no positive or negative inferences need be made from a student's failure to take the National Boards (obviously the same is true about the students in classes in which no grades were given). All missing values of National

Board scores were replaced by the average score of 500 points. This replacement would be inadequate for rank since those graduates applying for certification in internal medicine did significantly better in medical school than their classmates. Those students taking their first residency in internal medicine had an average rank of 527, and those who eventually took the written examination had an average rank of 541. Average among graduates is 500, and 541 is 0.41 standard deviations above average. To avoid biasing the estimated school effects at schools where some or all of the students had no grades, missing values of rank in class have been replaced with a value of 541.

HYPOTHESIS TESTING USING HIERARCHICAL STEPWISE REGRESSION

There are significant positive correlations between almost all pairs of variables considered in this report.[3] The largest correlations are between pairs of preadmission variables and pairs of medical school performance variables, but interperiod correlations are also large enough to cause some problems. Regressing the ABIM test score simultaneously against all the available independent variables would produce a good fit, but it would not indicate which variables had made the important contributions to the fit. Thus it would not be possible to distinguish between the effect of preadmission achievement and that of achievement during medical school, or the effect of the medical school itself or the postgraduate training that follows.

Careful selection of the independent variables should permit tests of the interesting hypotheses. Most of the results in this section are from a hierarchical stepwise regression done for the entire sample of physicians. At each step, selected variables, not necessarily those with the largest F-statistics, are allowed to enter the regression.[4] The order in which the independent variables are included is based on the hypothesized causal relationships among them. Although multiple linear regression is often used in prediction problems, the interest here is in testing a series of hypotheses. Therefore the values of coefficients are not as important as they would be to someone who is searching for the best prediction equation.

Step-by-step results of the regression are itemized in Table 1, which gives at each step the F-statistics and significance levels for which certain variables would enter at that point. The degrees of freedom associated with the preadmission variables differ from the rest since they are based on the subsample with no missing data. At the end of each step are the variables actually entered, with information on how they improve the fit. At Step 2, for example, rank and age were entered, even though several other variables would have been more significant than age.

One important hypothesis to test is that of the absence of school effects upon the ABIM test score. Since the National Board scores and postgraduate training characteristics are influenced by the medical schools, those are not suitable explanatory variables. National Board scores are correlated with the medical schools in much the same way as ABIM scores are. If these scores are included before the school effects are tested, some of the importance of the medical school would be lost. However, several variables were appropriate for inclusion before that regression step.

The first step included three dummy variables to allow for differences in scores among four different years of graduating classes.[5] The differences are highly significant, largely because the mean score for 1955 graduates was 40 points above the mean for the other years. Whatever the reasons for this, it seems unwise not to take it into account in the regression.[6]

RANK IN CLASS AND AGE AS EXPLANATORY VARIABLES

Two more explanatory variables were included before school effects were tested for. The first of these was rank in class. Since rank was standardized within each school in each year, it is not affected by the school in the same manner that National Board scores obviously are. Therefore, there is not the same objection to using rank as an explanatory variable as there would be to using the National Boards. Indeed, including rank in class controls for the differences in mean rank in class occurring at the different schools. For example, if students from all levels of one school attempted to be certified but only the top

Table 1

STEP-BY-STEP TABULATION OF REGRESSION RESULTS

(Except for preadmission variables, all items in the table refer to the basic stepwise regression of the first score on the written test for the full sample of 438 students. At each step the names of the variables which actually entered the regression are underlined.)

Variable	F	Degrees of Freedom	p-Level	Coefficient	Standard Error of Estimate	R^2
			STEP 1			
Year dummy variables	6.25	3, 434	<0.001		90.3	0.000[a]
			STEP 2			
Undergraduate GPA	24.51	1, 399	0.0			
Selectivity Index	7.51	1, 399	0.006			
Verbal MCAT	8.07	1, 399	0.005			
Quantitative MCAT	10.64	1, 399	0.001			
General Information MCAT	4.02	1, 399	0.045			
Science MCAT	21.98	1, 399	0.0			
Part II--Medicine	88.33	1, 433	0.0			
Part II--Mean	85.66	1, 433	0.0			
Part I--Mean	62.93	1, 433	0.0			
Specialized Internship	24.51	1, 433	0.0			
Major Teaching Hospital Internship	12.08	1, 433	<0.001			
M.T. Residency but not M.T. Internship	5.04	1, 433	0.025			
Waiting period	56.48	1, 433	0.0			
Rank	122.55	1, 433	0.0	0.463	79.6	0.221
Age	14.64	1, 432	<0.001	-8.00	78.6	0.246
			STEP 3			
School dummy variables	4.26	8, 424	<0.001		76.3	0.302

Table 1 (continued)

Variable	F	Degrees of Freedom	p-Level	Coef-ficient	Standard Error of Estimate	R^2
		STEP 4				
Undergraduate GPA	5.91	1, 389	0.015			
Selectivity Index	0.49	1, 389	(b)			
Verbal MCAT	1.07	1, 389	(b)			
Quantitative MCAT	1.23	1, 389	(b)			
General Information MCAT	0.20	1, 389	(b)			
Science MCAT	5.40	1, 389	0.02			
Part II--Mean	12.35	1, 423	<0.001			
Part I--Mean	4.52	1, 423	0.035			
Specialized Internship	1.47	1, 423	(b)			
M.T. Hospital Internship	0.56	1, 423	(b)			
M.T. Residency but not M.T. Internship	3.21	1, 423	0.075			
Waiting period	35.68	1, 423	0.0			
Part II--Medicine	17.28	1, 423	0.0	0.279	74.9	0.330
		STEP 5				
Part II--Mean	0.76	1, 422	(b)	0.066		
Part I--Mean	0.51	1, 422	(b)	0.058	74.9	0.332
		STEP 6				
Specialized Internship	0.74	1, 420	(b)	7.4		
M.T. Hospital Internship	0.37	1, 420	(b)	-3.3		
M.T. Residency but not M.T. Internship	2.83	1, 420	0.09	-24.5	74.9	0.337
		STEP 7				
Waiting period (in years)	39.98	1, 417	0.0	-16.3	71.6	0.395

[a]All values of R^2 are calculated relative to the mean squared residual at the completion of Step 1. Before this step the standard error was 91.9.

[b]The variable was not significant at the 0.20 level.

ones from a weaker one did, no difference might be observed between the schools unless rank were used as a covariate. As expected, rank in class is a powerful explanatory variable. Its regression coefficent after the step introducing school dummies is 0.466 (rank has approximately the same scale as the written examination).

Also included at this step was the age at graduation of the students (values exceeding 34 years were lowered to that level to reduce the effect of any one person's score). There is no reason to expect that age would relate to school in any way that would bias the school effects. Delays in reaching medical school were correlated with poor performance on the ABIM examination. The regression coefficient on age was -8.0 points per year (standard error = 2.1).

TEST FOR SCHOOL EFFECTS

To test one of the most important hypotheses under consideration, whether there are differences among the schools, school effects may be estimated by putting dummy variables for each medical school into the regression. These estimated coefficients, normalized to sum to zero, are given for overall and single year regressions in Table 2. Under the hypothesis that the medical school makes no difference in scores, the expected value of each of these coefficients would be zero. In the regression covering all four years, this hypothesis may be rejected with near certainty ($F = 4.26$). Regressions for the separate years produce similar coefficients, although the school differences are significant only twice. If terms are introduced to allow the school coefficients to change linearly over time, the one for school 4 is significant at the 0.02 level, but no others are significiant at the 0.1 level. No overall test of the hypothesis that the school coefficients do not change over time is even marginally significant. The coefficients for the class of 1955 could not be published (see notes to Table 2). Those numbers exactly follow the sign pattern of the overall coefficients, although they are somewhat more dispersed.

No one would dispute that the schools with the best reputations are in a position to choose from candidates with the best credentials. There are unquestionable differences in the quality of students admitted

Table 2

ESTIMATED SCHOOL COEFFICIENTS OVER TIME[a]

Medical School Number	All Years[b]		1955	1960	1965	1969
1	27.5	(33.6)		20.9	17.5	39.0
2	27.5	(17.5)		16.6	50.5	4.6
3	18.8	(14.6)		6.8	10.9	51.3
4	9.4	(0.0)		19.2	-14.0	-8.1
5	5.8	(9.8)		-1.3	15.0	-25.2
6	-16.8	(-9.0)		-27.2	-27.8	-0.4
7	-17.6	(-18.6)		-14.1	-1.5	-19.3
8	-27.0	(-23.5)		0.5	-55.6	-8.3
9	-27.6	(-24.5)		-21.3	4.8	-33.4
Number of observations	438	404	79	116	130	113
F-statistics[c]	4.26	2.19	2.17	0.95	1.38	2.63
Significance level	0.001	0.025	0.05	--	--	0.015
Pooled estimate for standard error[d]	10.9	13.8	26.2	18.2	25.7	17.5

[a]One of the medical schools did not graduate a class in 1955. To preserve the anonymity of the results for that school, the school coefficients for 1955 are not given.

[b]The numbers in parentheses are the estimated school effects after preadmission variables as well as age and rank in class are controlled for.

[c]The F-statistics test the hypothesis that there are no differences among the schools. They are based on 7 or 8 degrees of freedom in the numerator and a minimum of 69 in the denominator.

[d]Although the standard errors of the estimated coefficients are somewhat different for each of the schools, a pooled standard error was computed for each regression by the following method: If $\{\hat{\beta}_i\}$ are the estimated coefficients and σ is the common standard error, then $\sum_i \left(\frac{\hat{\beta}_i}{\sigma}\right)^2 = (n-1)F$, where n is the number of schools. Thus $\hat{\sigma}^2 = \frac{\Sigma \hat{\beta}_i^2}{(n-1)F}$.

at the schools. Whether these differences exist and are important within the rather select group who finally take the ABIM examination is in doubt. To test this, the six preadmission variables were introduced into the regression (Step 4). Either the undergraduate grade point average or science MCAT would be significant at the 0.02 level, but the rest are not at all significant. These variables increase the goodness of fit of the regression only very slightly, but they do influence the estimated school effects. The revised school coefficients are listed in parentheses in Table 2, and they tend to move toward zero. As a set, the nine school coeffcients are still significant at the 0.025 level. There were definite, substantial differences in the levels of performance of graduates of the nine medical schools, and preadmission variables account for much of these differences. Whether the remaining differences are due to the schools or simply to an inability to measure preadmission characteristics perfectly remains unanswered.

For this sample of nine medical schools, there is a strong relationship between the amount of research being done at the institution and the school effect on ABIM test scores. In selecting the academic health centers for this study, the Rand study group developed a factor measuring the research intensity at each of the medical schools. The factor measures the amount of NIH research, the amount of other research, and the number of residents and interns at affiliated teaching hospitals. The research factor is highly correlated with the quality of incoming students. The correlation of research with mean science MCAT is 0.95 and with mean undergraduate GPA is 0.70 (for all 1955, 1960, 1965, and 1969 graduates).[7] Correlation of the research intensity factor with the school effects on the ABIM test is 0.72; after individually controlling for premedical school achievement, it is 0.65.

INTRODUCTION OF NATIONAL BOARD SCORES

Certain of the National Board scores correlate with the written ABIM test scores as much as or more than rank in class does. Since the medical schools are likely to affect the scores of their students on the National Boards, rank was used as an explanatory variable first. After the step

checking for school effects, it is now desirable to see how much the fit can be improved. At Step 4, the F-statistics for any one of them to enter the regression are 17.3 for the medicine section of Part II, 12.4 for the mean on Part II, and 4.5 for the mean on Part I. After the medicine score has been entered, neither of the others is even close to being significant. These results are not representative of the importance of the National Boards in the regression, however, since only 49 percent of the sample took Part I and 54 percent took Part II.

The following are the results of a regression restricted to the six medical schools where most of the students took the National Boards, and at those schools, to those students who took Part II. Rank in class is available for everyone in the sample, and most of the students had taken Part I so maximum information is available for the regression. The initial steps of the regression are similar to those presented on the preceding pages with one exception. The six schools in this regression are, perhaps, the most homogeneous group in the sample. The initial standard error after controlling for year of graduation is only 81.8, compared with 90.3 for the entire sample at the same point. The school coefficients are still significant at the 0.015 level ($F = 2.91$). The big change is that Part II of the National Boards is now very important in improving the fit of the regression equation. The three test scores listed above now have F-statistics to enter of 42.4, 35.5, and 8.2. After the medicine section of Part II has been entered, the mean on Part II is still significant ($F = 5.1$), but the mean on Part I is not ($F = 1.2$). The coefficients of both Part II scores are higher than the coefficient of rank in class. The standard error of the regression is reduced to 61.2 for a value of $R^2 = 0.464$.

POSTGRADUATE TRAINING VARIABLES

No significant results were found concerning the three postgraduate training variables in any regression. Two points explain why this is the case. First, almost all medical school graduates who subsequently seek certification have both internships and residencies at major teaching hospitals and, recently, specialized internships also. Thus,

these distinctions have little explanatory power. Secondly, since the predictive power of the undergraduate and medical school variables is high, it seems likely that the marginal effect of better training variables probably would be small.

RELATIONSHIP OF SCORE WITH DELAY IN TAKING TEST

Before taking the written examination, the candidate must satisfy the Board's postgraduate training requirements. Therefore, the earliest that a 1955 or 1960 graduate could attempt the written examination was six years after graduation. Such steps as allowing the substitution of a straight medicine internship for one year of residency have been taken to shorten this period, so that 1965 graduates had a minimum wait of four years and 1969 graduates only three years. For any number of reasons, many candidates wait an additional one, two, three, or more years before taking the exam. Table 3 shows the relationship between year of graduation and when the test was taken.

How long after graduation the examination is first attempted is related to the score on it; knowing nothing more about a student than how much longer than the minimum time he waited explains 11.5 percent of the variance in scores. Those students who take the test immediately do much better than those who wait a number of years. Three factors may contribute to this phenomenon. First, the best students are able to find places immediately in preferred straight medicine internships and medicine residencies. Some students with unimpressive records may need to prove themselves before being admitted to the programs required for admission to the written examination. This reason is supported by the fact that how well a student did in medical school is highly correlated with how soon he takes the exam. A second factor is that some weaker students may spend extra time in training or even several years in practice before having the confidence to apply for certification. Finally, on such a test, freshness of knowledge gained during medical school and early postgraduate training may simply outweigh the value of experience gained during practice.

Table 3

TABULATION OF THE FIRST ATTEMPT AT
THE WRITTEN EXAMINATION BY YEAR OF GRADUATION

	Year of Graduation			
	1955	1960	1965	1969
Students with first residencies in medicine	111	174	172	204
Had not taken written exam by 1973	32	58	42	91
	Years after Graduation			
	6	6	4	3
Took exam immediately	37	70	37	59
1 year delay	17	22	43	54
2 year delay	9	15	not given	
3 year delay	5	4	43	
4 year delay	4	4	7	
5-7 year delay	4	1		
8-12 year delay	3			

The importance of the waiting period decreases as more is known about the physician's background (compare Steps 2 and 7), but even at the last step of the regression its coefficient is -16.3 points per year with a standard error of 2.6. The coefficient is highly significant in separate regressions for graduates of 1955, 1960, and 1965. The 1969 regression only compares physicians who took the test in 1972 with those who took it in 1973. To take the examination in the summer of 1972, a 1969 graduate necessarily had taken a medicine internship and exactly two years of an internal medicine residency. For 1969, the timing effect is positive and nonsignificant. In the regression with the most complete information, that one limited to physicians who had taken Part II, the coefficient falls to -10.2 points per year with a standard error of 3.4. This is still significant at the 0.003 level.

The fact that the timing effect is quite significant even in the best regression indicates that there is considerable systematic variance in abilities for which the usual indicators cannot account. These abilities, however, are discernible in some way to the candidates themselves and to the medicine program directors who influence when the examination is taken.

FOOTNOTES TO SECTION II

[1]Also available was the science grade point average, but there were missing values for over one-third of the students. When both are available, the two are nearly identical in predictive ability.

[2]Postgraduate training information was obtained from the AMA Masterfile of Physicians as of December 31, 1972.

[3]In general, there is a *negative* correlation between age at graduation and any of the other variables, including the written examination score.

[4]The procedure used here differs from an ordinary stepwise regression, which is an algorithm for searching for the best subset of independent variables to include in the regression. At the first step of

an ordinary stepwise regression, the variable with the largest F-statistic is entered. At later steps, an independent variable is not entered until it possesses the largest "F to enter."

[5]There are three separate dummy variables for the graduating class years 1960, 1965, and 1969. Each takes a value of 1 if the graduate is from that year; otherwise the value of zero is assigned. If the graduate is from the class of 1955, the values for all three dummy variables are zero, and the 1955 effect is reflected in the regression constant term.

[6]The use of a dummy variable affects the constant term of the linear equation but not the slope, which reflects a continuous pattern of association between the multiple independent variables and the single dependent variable, the ABIM test score.

[7]With only nine pairs of observations, the correlation coefficient is quite unstable. Significance at the 0.05 level requires a coefficient of 0.67, and significance at the 0.01 level requires one of 0.80.

III. MEASUREMENT ERROR IN THE ABIM WRITTEN EXAMINATION

As was discussed in the first section, the ABIM written examination cannot test all the qualities that combine to make a good internist. It *is* designed to test the examinee's comprehension of the large body of knowledge required by a good internist. Because the body of knowledge is so large, even such a long test can cover only a small fraction of it. Many of the questions are changed each year so that a person's score may fluctuate depending on the edition of the examination he is given. Because of this fluctuation, the "true" score becomes an important notion. It is defined to be the expected value of the observed score over the distribution of possible outcomes.[1] The difference of the true score from the observed score is defined to be the "error" score. Although it need not be the case, one would hope that the true score on a test would be very highly correlated with the skills it is intended to measure. One would also like the variance in error scores--the "measurement" error--to be small compared with the variance in true scores across the population.

Knowing the size of the measurement error is important in judging the quality of fit of the regressions in the last section. Since the error score is independent of the various prediction variables that appear in the regressions, a certain amount of the variance in observed scores is a priori "unexplainable." Knowing the size of the measurement error would allow us to estimate how much of the variance in the true scores is explained by the regressions, even though the true scores are unknown.

An estimate of the measurement error of the test may be obtained through an item analysis of the test questions. For examinees at a particular level of ability, the variance due to each item on the test is known. The sum of these variances is the lower bound for the measurement error in the raw score, which is easily transformed to the standardized scale. The ABIM provided estimated variances obtained in a manner similar to this for the past ten years, ranging from approximately 500-1,000, the square roots of these ranging from about 22-33 points.

This method does not account for the presence of series of questions about case histories, or a given drug or procedure. Questions in series have higher within-series correlations than with items elsewhere in the tests. That is particularly true within patient management problems since the correct answer to one item may help in answering later ones. This phenomenon increases the actual measurement error in the standardized score. Intuitively, a series of related questions adds more variance to the raw scores than a series of unrelated ones but actually covers less of the knowledge in the subject area. For these reasons, another method to estimate the average measurement error over the 1961-73 period would be desirable. Substantially larger errors would be expected in the one-day editions of the test given before 1972.

An alternate estimate is available using the data supplied by the Board. Let σ^2 denote the variance in test scores due to measurement error. The simplifying assumption is made that σ^2 is the same for all examinees in all years. Besides the item-by-item variance, σ^2 allows for any random factors affecting the scores--headaches, family problems, good days, bad days, etc. The scores of 84 physicians, those who took the written examination more than once, are used to estimate σ^2. A few standard assumptions permit derivation of an expression for the conditional distribution of an individual's second score, given his first score. The variance of this distribution, which can be estimated using linear regression, is a function of σ^2 and σ_Δ^2 (the variance of the real change in ability between the first and second tests). If $\sigma_\Delta^2 = 0$ (i.e., everyone's knowledge increases or decreases by the same amount), then $\hat{\sigma}^2 = 2440$ with $SE(\hat{\sigma}^2) = 460$. Thus $\hat{\sigma} = 49.4$. Of course, if $\sigma_\Delta^2 > 0$, then the estimate of σ^2 should be revised downward. The model and estimation techniques used to obtain this result are described fully in the appendix.

The two methods described here provide both a lower and an upper bound for the measurement error. The first one is suspected to be an underestimate and the latter an overestimate, but one cannot determine the precise bias associated with either. The average measurement error for the period is very likely in the range of 30-45 points, but the

present measurement error for the test might well be below that range.

FOOTNOTE TO SECTION III

[1]This is the definition used by F. M. Lord and M. R. Novick in *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Mass., 1968. It turns out to be equivalent to the often stated definition: The true score is the value to which the sample mean score would converge if the test could be replicated a large number of times.

IV. CONCLUSIONS

A number of least squares hierarchical stepwise regressions have been performed to predict the score on the first taking of the ABIM written examination. The explanatory variables include preadmission variables, medical school performance variables (as well as the school itself), and postgraduate training dummy variables. A step-by-step record of the basic regression involving 438 physicians is given in Table 1.

The emphasis of the regression work is in discovering how much various factors contribute to differences in scores on the ABIM written examination. This helps formulate answers to such questions as: How well can we predict how an incoming medical student would do on the exam? Does the medical school attended make a difference in the scores? If so, how large is the contribution? What other factors affect the scores? And how much of the variance is explained by previous performance? To the extent that the ABIM examination measures the quality of care provided by the physicians who take it, it is possible to make cautious extrapolations of the relationship of performance on the test to performance in the field.

The most important question of policy interest concerns how different medical schools affect the examination scores. Without controlling for differences in the abilities of incoming students, "school effects" were estimated for the nine medical schools in the sample.[1] The school effect is the number of points that the score of a student from any level of his class--top, middle, or bottom--would be expected to change merely because he is from that particular school. The effects are quite significantly different from zero, covering a range of 55 points for these schools. Still the rank within a class is much more important than the class itself. In terms of variance explained, the medical school accounts for 25 percent as much as the single variable rank in class does. The effects seemed to be quite stable. At only one school was there evidence of a linear trend over time. The estimated school effects exhibit a surprising correlation with a composite measure

of research activity at the institutions, although it is unwise to draw too strong a conclusion from so small a sample of schools.

As well as differences in the scores of the graduates, there are differences in the abilities of the entering classes at the schools. When preadmission variables (undergraduate grades, selectivity of undergraduate college, MCATs) are included, most of the estimated school effects move toward zero. The school variables are still significant, but only at the 0.025 level. The correlation with research activity is still fairly high, indicating that there were definite differences in performance on the ABIM examination by graduates of the various medical schools. Part of that is certainly due to differences in the incoming students. Indeed, if it were possible to measure preadmission performance better (for example, with letters of recommendation), there might be no difference at all.

Other findings of the study are the following:

1. There are significant positive correlations between the score on the written examination and *each* of the variables considered--the preadmission variables, age, rank in class, National Board scores, and the postgraduate training dummy variables. There are also significant positive correlations between most pairs of these variables.[2]

2. Rank in class and the National Board scores are the best predictors of success on the ABIM examination. The best pairwise correlations with the score on the written examination are with rank in class (.5162), mean on Part I (.5300), medicine test on Part II (.6145), and mean on Part II (.6145). Since little more than half the sample took the National Boards, rank in class is a more useful predictor than any of the Board scores. The correlations with other individual tests are also high but would provide no marginal help. The mean on Part I, which is generally taken in the second year of medical school, is only useful in the absence of information about Part II, which would be taken in the fourth year. Of the other two, the medicine section of Part II is slightly the better predictor, but any two of the three explain about as much as all three would.

3. The ability to predict scores based on these variables available before entrance to medical school is only fair ($R^2 = 0.136$).[3] Even

though this is significant, it does not imply any success for admissions committees in selecting candidates who will do well on the ABIM examination since these committees have little idea *which* students will choose to specialize in internal medicine. The effects of these variables are dwarfed by the rank in class of the student, which alone provides an $R^2 = 0.221$. After rank has been included, only age, undergraduate GPA, and the science MCAT are significant.

4. After controlling for school, age, rank and National Board scores, none of the postgraduate training variables are significant. Unfortunately no criteria for grading hospitals or training programs other than an admittedly crude major teaching hospital distinction were available. Thus there is no good basis in these data for judging the marginal importance of postgraduate training on the examination score. More detailed information on the postgraduate period would be necessary to separate these effects, if they exist, from those already measured.

5. Separate regressions for each of the four years provide generally consistent estimates of the regression coefficients. The only exception is that rank in class was a better predictor for scores of 1955 graduates than those of later years.

6. The regression analysis shows that a large fraction of the variance in examination scores may be accounted for by differences in previous achievement. Age, rank in class, medical school, and the medicine section of the National Boards were used to reduce the standard error of the estimate for the entire sample from 90.3 to 74.9 ($R^2 = 0.330$).[4] This reduction is limited in large part by the absence of National Board scores for many students. Using the same variables plus the mean on Part II (only for those students who had taken Part II), the standard error was reduced from 81.8 before the regression to 61.2 ($R^2 = 0.464$). Since there is a measurement error inherent to the test, each of these regressions accounts for a larger part of the "explainable" variance than the value of R^2 indicates. I estimate that the average standard error of measurement during the time period in this study is in the range of 30-45 points. Thus about 11 to 25 percent of the variance is "unexplainable" by these regressions.

7. The minimum time from graduation until a physician could first take the written examination has decreased considerably. For 1955 and 1960 graduates it was six years; now it is three.[5] How soon after the minimum wait the candidate attempts the test is related to how well he would be expected to do. Those physicians who take the test immediately do much better than those who wait a number of years. Even after controlling for medical school performance in the regression that included only students who had taken Part. II of the National Boards, it is still very significant. The estimated difference in scores is -10.2 points per year delay with a standard error of 3.4. The significance of this variable indicates that there is additional systematic variance recognizable by the physician, his advisors, and program directors (the group of people who influence when the examination is first taken) for which the other variables cannot account.

Some proportion of the variance in scores is due to random fluctuations occurring at the time of the test. This will always be the case for any test, and all that can be done is to try to minimize the random factor, for example, by including more questions. What is most important is that the systematic variance--not due to random factors--measures knowledge that is required of an internist.

The determination of what knowledge an internist requires is, of course, the responsibility of those who prepare the examination. What this report has addressed are the characteristics of examinees and their training that explain systematic variance in test performance. The more complete the explanation of systematic variance, the more safely it may be assumed that differences in scores reflect differences in knowledge levels of examinees. This is consistent with the regression results that show rank in class, scores on Part II of the National Boards, and the other variables accounting for a majority, perhaps a large majority, of the systematic variance in test scores.

FOOTNOTES TO SECTION IV

[1]See Table 2. The effects are normalized to sum to zero in each column.

[2]In general there is a *negative* correlation between age and any of the other variables, including the score on the written examination.

[3]That is, a regression with these variables accounts for 13.6 percent of the variance in scores.

[4]90.3 differs from the sample standard deviation of 91.9 since it measures only deviations within each graduating class. See Step 1 in Table 1.

[5]See Table 3.

Appendix

USING REPEATED TAKINGS TO ESTIMATE THE MEASUREMENT ERROR OF THE WRITTEN EXAMINATION

In Section III there was a discussion on the importance of measurement error in the written examination scores. This appendix develops a model that permits estimation of that quantity using the scores of physicians who repeated the examination. The following statistics were compiled for the entire sample.

Taking Number	Number of Observations	Mean Score	Standard Deviation
1	438	506.5	91.9
2	84	450.7	74.3
3-5	26	419.6	58.2
All	548	493.8	91.7

Since the best students pass on their first attempts, the mean score on the second taking is substantially lower. A similar though smaller decline occurs between the second and third takings. This sample of students is very similar to the reference group used to standardize the scores, so the mean on taking the first test should be near 500 and the standard deviation of those scores should be near 100. The mean score is not significantly different from 500. The standard deviation is significantly below 100, but this may be due to variations in the standardization procedure. Scores for the 84 physicians who repeated the examination are listed in Table A.1.

DEVELOPMENT OF THE MODEL

Let X_{1i}, X_{2i} denote the ith person's scores the first and second times he takes the written examination; X_{2i} would generally be observed only if X_{1i} is less than the minimum passing score.[1] Then it is assumed that

$$X_{1i}|\mu_i \sim N(\mu_i, \sigma^2) \qquad (1)$$

Table A.1

FIRST AND SECOND SCORES OF 84 PHYSICIANS WHO REPEATED THE ABIM WRITTEN EXAMINATION

First	Second	First	Second
280	390	440	420
290	370	440	480
300	390	440	480
310	380	450	320
320	380	450	470
330	370	450	470
330	390	450	480
340	320	460	380
350	290	460	450
350	310	460	450
350	370	460	460
350	430	460	500
350	440	460	590
360	470	470	370
370	260	470	380
370	410	470	390
380	430	470	430
380	480	470	470
390	350	470	490
400	310	470	510
400	350	470	550
400	420	470	580
400	420	480	450
400	460	480	480
400	470	480	590
410	470	490	410
410	480	490	510
410	520	490	540
410	530	500	430
410	580	500	490
420	380	500	520
420	400	500	550
420	420	500	560
420	420	500	590
420	440	510	470
420	440	520	530
420	440	530	450
420	560	530	480
430	440	530	480
430	490	550	460
430	530	570	550
430	580	590	420

where

$$\mu_i \sim N(\mu, \sigma_\mu^2) \ . \tag{2}$$

Then μ_i is the score that corresponds to the ith person's true ability at the time of his first attempt, and σ^2 is the variance due to measurement error. It is assumed that σ^2 is the same for everyone. Then the distribution of X_1, the score on the first attempt by a randomly selected student, is

$$X_1 \sim N(\mu, \sigma^2 + \sigma_\mu^2) \ .$$

For the nine medical schools studied, μ and $(\sigma^2 + \sigma_\mu^2)^{1/2}$ may be estimated by the sample mean and sample standard deviation for the first taking; these are 506.5 and 91.9 respectively.

The next assumption is that

$$X_{2i}|\mu_i, \Delta_i \sim N(\mu_i + \Delta_i, \sigma^2) \ , \tag{3}$$

where Δ_i, the change in true ability, is independent of μ_i and X_{1i} and is distributed as

$$\Delta_i \sim N(\Delta, \sigma_\Delta^2) \ . \tag{4}$$

The assumption that μ_i and Δ_i are independent is a reasonable one even though students with low values of μ_i have "more to learn" than ones with higher values of μ_i. If so, then the assumption of independence between X_{1i} and Δ_i should also be valid.

By studying the joint distribution of X_{1i} and X_{2i} for i such that X_{2i} is observed, estimates may be developed ot the mean improvement in ability between takings (Δ) and of the size of the measurement error (σ^2). For the sample of 84 cases, the improvement in scores ($X_{1i} - X_{2i}$) had a mean of 17.6 and sample standard deviation of 69.7. Unfortunately, the sample is not randomly drawn from the joint distribution of (X_{1i}, X_{2i}) since whether X_{2i} is observed depends on X_{1i}. With very few exceptions, those students passing on their first attempts do

not retake the test. Students who score just below the passing level are almost certain to retake the examination, and those who score much lower are likely to be discouraged from making further attempts. Since $(X_{2i} - X_{1i})$ depends on X_{1i}, its sample mean, 17.6, would be a very bad estimate of Δ.

The problem described above can be alleviated by analyzing the distribution of X_{2i} conditioned on the value of X_{1i}. As a first step, the distribution of μ_i, conditioned on X_{1i}, is given by

$$\mu_i | X_{1i} \sim N\left(\left(\frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}\right) \mu + \left(\frac{\sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}\right) X_{1i} , \frac{\sigma^2 \sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}\right). \qquad (5)$$

Then by (3) and (5),

$$X_{2i} | X_{1i}, \Delta_i \sim N\left(\left(\frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}\right) \mu + \left(\frac{\sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}\right) X_{1i} + \Delta_i, \sigma^2 + \frac{\sigma^2 \sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}\right),$$

and from (4)

$$X_{2i} | X_{1i} \sim N\left(\left(\frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}\right) \mu + \left(\frac{\sigma_\mu^2}{\sigma^2 + \sigma_\mu^2}\right) X_{1i} + \Delta, \sigma^2 + \frac{\sigma^2 \sigma_\mu^2}{\sigma^2 + \sigma_\mu^2} + \sigma_\Delta^2\right). \qquad (6)$$

This last equation provides methods to estimate the unknown parameters of the model. Before the estimation procedures are described, it is desirable to examine the assumptions that have been made.

1. The random variables in (1)-(4) have normal distributions.
2. The value of σ^2 is independent of μ_i.
3. The random variable Δ_i is independent of μ_i and X_{1i}.
4. Given X_{1i}, whether X_{2i} is observed is independent of μ_i.

The first two are standard assumptions and the third one was discussed earlier. The last assumption implies that the conditional distribution of μ_i, given X_{1i} *and* that X_{2i} is observed, is the same as

that in (5). There will be a breakdown here if physicians who do better than expected and obtain a certain failing score have a lower probability of retaking the test than their counterparts who do worse than expected in obtaining the same score. This possibility will be discussed later.

PARAMETER ESTIMATION METHODS

From (6) the expected value of X_{2i} is a linear function of X_{1i}, so that the model

$$X_{2i} = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \qquad (7)$$

holds. Suppose the error term, ε_i, has variance η^2. Then by (6)

$$\eta^2 = \sigma^2 \left(2 - \frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}\right) + \sigma_\Delta^2 .$$

The mean squared residual in the regression of X_{2i} on X_{1i} is a minimum variance unbiased estimate for η^2. If σ_Δ^2 is specified, then it is possible to obtain a good estimate of σ^2 by substituting the mean square residual for η^2. Setting $\sigma_\Delta^2 = 0$ gives an estimate for the upper bound of σ^2. In that case the quadratic formula yields

$$\sigma^2 = \left(1 \pm \left(1 - \frac{\eta^2}{\sigma^2 + \sigma_\mu^2}\right)^{1/2}\right)(\sigma^2 + \sigma_\mu^2) ,$$

and since $\sigma^2 < \sigma^2 + \sigma_\mu^2$,

$$\sigma^2 = \left(1 - \left(1 - \frac{\eta^2}{\sigma^2 + \sigma_\mu^2}\right)^{1/2}\right)(\sigma^2 + \sigma_\mu^2) . \qquad (8)$$

The mean square for the regression equals 4176 with a standard error for the mean square of $(4176)(2/82)^{1/2}$. Substituting the mean square for η^2 in (8) gives

$$\hat{\sigma}^2 = 2440 .$$

The standard error of the estimate is

$$SE\ (\hat{\sigma}^2) = 460 .^{2}$$

For values of $\sigma_\Delta^2 > 0$, $\hat{\sigma}^2$ would decrease at about 60 to 70 percent of the rate at which σ_Δ^2 increased.

There is another method to estimate σ^2 using the regression in (7). From (6) and (7),

$$\beta_1 = 1 - \frac{\sigma^2}{\sigma^2 + \sigma_\mu^2} ,$$

so that

$$\sigma^2 = (\sigma^2 + \sigma_\mu^2)(1 - \beta_1) .$$

Thus σ^2 is easily estimated from the estimates of $(\sigma^2 + \sigma_\mu^2)$ and β_1. From the regression $\hat{\beta}_1 = 0.579$, so

$$\hat{\sigma}^2 = (91.9)^2\ (0.421) = 3550 .$$

The estimated standard error of this estimate is

$$SE\ (\hat{\sigma}^2) = 960 .^{3}$$

Additionally, from (6)

$$\beta_0 = \left(\frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}\right) \mu + \Delta = (1 - \beta_1)\ \mu + \Delta .$$

Thus an unbiased estimate for Δ is

$$\hat{\Delta} = \hat{\beta}_0 - (1 - \hat{\beta}_1)\ \hat{\mu} .$$

Using $\hat{\mu} = 506$ gives $\hat{\Delta} = -13.1$, which is not significantly different from zero.

The first estimate of σ^2 is much more precise than the second one since its standard error is less than half. The second one is also much more sensitive to failures in the third and fourth assumptions. A check was made for a nonlinear dependence of X_{2i} on X_{1i}, but none was found. Nevertheless, if only the best of the physicians who had scored very low the first time retook the test, then the dependence of X_{2i} on X_{1i} would be smaller than suggested by (6). As a result, the second estimate of σ^2 would be too high. Conversely, the former estimate would be affected little, if any. Of course, it is still necessary to know σ_Δ^2 for this estimate to be accurate.

There is a considerable difference between the two estimates, which may be due to the problems discussed above or may be purely random. In any case, the former estimate, $\hat{\sigma}^2 = 2440$ ($\hat{\sigma} = 49.4$), is the more reliable. Given that σ_Δ^2 is certainly positive, is it reasonable to expect that σ^2 is actually smaller than that estimate?

FOOTNOTES TO APPENDIX

[1]Our sample does contain records on seven students who passed the first attempt and later took the test again as an alternative to taking the oral examination during its phasing out period.

[2]The estimated value of SE $(\hat{\sigma}^2)$ is obtained from the approximate formula

$$\text{SE}\ (\hat{\sigma}^2)^2 = \left(\frac{\partial\, \sigma^2}{\partial\, \eta^2}\right)^2 \text{SE}\ (\hat{\eta}^2)^2 + \left(\frac{\partial\, \sigma^2}{\partial\, (\sigma^2 + \sigma_\mu^2)}\right)^2 \text{SE}\ \widehat{(\sigma^2 + \sigma_\mu^2)}^2$$

evaluated at the sample estimates. The derivatives are

$$\frac{\partial\, \sigma^2}{\partial\, \eta^2} = \frac{1}{2\left(1 - \dfrac{\eta^2}{\sigma^2 + \sigma_\mu^2}\right)^{1/2}}$$

and

$$\frac{\partial \sigma^2}{\partial(\sigma^2 + \sigma_\mu{}^2)} = 1 - \left(1 - \frac{\eta^2}{\sigma^2 + \sigma_\mu{}^2}\right)^{1/2} - \frac{\eta^2}{2(\sigma^2 + \sigma_\mu{}^2)\left(1 - \frac{\eta^2}{\sigma^2 + \sigma_\mu{}^2}\right)^{1/2}}$$

thus

$$SE(\hat{\sigma}^2)^2 = (0.703)^2\ (4176)^2\ (2/82) + (-0.0587)^2\ (91.4)^4\ (2/437)$$
$$= 210{,}335 + 1{,}125 = 211{,}460,$$

so that

$$SE(\hat{\sigma}^2) = 460\ .$$

[3]By the approximate standard error formula used above

$$SE(\hat{\sigma}^2)^2 = \widehat{(\sigma^2 + \sigma_\mu{}^2)}^2\ SE(\hat{\beta}_1)^2 + (1 - \hat{\beta}_1)^2\ \widehat{SE(\sigma^2 + \sigma_\mu{}^2)}^2$$

$$= (91.4)^4\ (0.110)^2 + (0.421)^2\ (2/437)\ (91.9)^4$$

$$= 863{,}000 + 59{,}000 = 922{,}000$$

so that

$$SE(\hat{\sigma}^2) = 960\ .$$